Richard Stefanik

Spring 2018 Research Summary

Given data from a selection of Notre Dame students, I experimented with different modeling techniques to better understand sleep patterns. The given data included sleep data, activity data, mobile phone screen usage data, mobile phone audio usage data, weather data, and survey data. The sleep data and activity data came from Fitbit Charge HR devices the subjects wore. The screen usage data and audio usage data came from tracking the subjects' mobile phones. The weather data was simply the weather recorded for the South Bend area over the given time period, and survey data came from a Pittsburgh Sleep Quality Index (PSQI) survey the subjects periodically filled out. The main approach for this research project was to first organize the data using Matlab, then generate different types of models using RStudio, then return this information to Matlab for analysis.

The first major step was to organize all the data into a single csv file that could be used by RStudio to generate the different models. This process involved transforming some of the data files into more readable csv files, in order to more easily combine all the data later on. For instance, the original screen data file only had UNIX timestamps and whether or not the phone had been locked or unlocked at that timestamp. This data had to be further processed in order to scan for anomalies (i.e. the data says that the user unlocked the phone twice without ever locking it) and outliers, determine a given screen use session, and put dates to screen session. (This file was called DetailedScreenSessions_#.csv, where # is the subject number.) Similar processes were carried out for all the raw data files. For all the processing of the original raw data files, keeping track of the date was important, since the ultimate goal of all these processing was to combine all of this data into a single csv file, synchronized by day.

With all the data processed, the Matlab script processMetaFile.m was able to be run, which accomplished the task of combining all of this data. This file is called DataMetaFile.csv. (The Notre Dame Academic calendar was also analyzed in order to give more data for this meta file. For example, was the given day during the a break, exam week, reading day, etc.) In order to better understand the coefficients of the model that would be produced in RStudio, all the values that would be independent variables in the models were normalized and stored in a new file called NormalizedMetaFile.csv. Finally, the last step before moving over to RStudio was to separate all this data into two different groups, one for training and one for validation. The training data was 75% of the original data, and the validation data was the other 25% of the data. There was no overlap between the training and validation data. (Training and validation groups were made for the entire population and for each subject, as some models in RStudio would be run for the entire population while others would be run for single subjects.)

Moving into RStudio, two different modeling techniques were utilized: the General Linear Model (GLM), which is designed for single subjects, and the Generalized Linear Mixed Model (GLMM), which is designed for a population, and accounts for the differences in subjects. A GLM was also run for the entire population, treating all subjects as one, even though this modeling technique is not designed to be used in this way. In total three different types of models were made: GLM for each subject individually, GLMM for the entire population, and GLM for the entire population. These models were all run to determine sleep duration based on the independent variables from the NormalizedMetaFile.csv. All three types of models were first run with all the possible factors (independent variables), and then only run with factors the first model deemed significant. A Chi-Squared test was carried out between the models with all factors and models with only significant factors where applicable.

With the models generated and the appropriate data carried over from RStudio, Matlab was used to plot the model's prediction for sleep duration vs. actual sleep duration, now using the validation data. In visually analyzing the graphs, it was quickly noticed that a few outliers in these plots were likely greatly affected the r-values of these plots. Because of this, a density based scan, or DB scan, was employed in an attempt to remove these outliers and better understand the performance of the model. The density based scan accepts a minimum number of points and a distance value. A given point is selected, if that point does not have the minimum number of points within the given distance value, it is marked as an outlier. If that point does have the minimum number of points within the given distance value, then that point and all other points are not outliers. Using this approach the r-values did improve, but it soon became evident that using a constant minimum number of points and distance value across all plots was not effective. Instead, for each plot, the DB scan was run with a large number of different parameters in order to determine which yielded the highest r-value. These plots can be seen in the Matlab scripts processGLMPlots.m and processPopulations.m. The data for all these plots are organized into two csv files, GLM_AllSubjects_Data.csv, which is for the GLM models run on individuals, and Population_Data.csv, which is for the GLM and GLMM models run on the entire population. These files included data for plots with all factors and only significant factors, and with and without outlier removal.

Nap duration was consistently negatively correlated with sleep duration, meaning if someone took a nap that day, they would get less sleep that night. Other factors like weather seemed to have little to no effect. No conclusive evidence evidence could be drawn on the effects of screen and audio usage on the whole, but for the most part, audio usage thirty minutes before bed was positively correlated, meaning that those who listened to something on their phone before bed slept longer that night. This perhaps could be subjects listening to white noise to aid them in falling asleep. As expected, the GLM for individuals greatly outperformed the GLM and GLMM for the entire population. Since sleep habits can vary so much from person to person, it makes sense that having an individualized model is preferable to a model for the entire population.

In my opinion, it would be best to focus on the GLM for individuals and work to improve these even further. I do see another route of attempting to improve the population models, but since their r-values were only around .4, while the r-values of the GLM individual models were around .65, I think the first choice is preferable. More subjects could certainly be analyzed, as I only worked with the data for around 25 subjects. Perhaps having more subjects could aid the population models. More factors could be added to the model, but these factors have to be thought about carefully. In the models run, an interaction factor for screen usage thirty minutes before bed and audio usage thirty minutes before bed was included, but only because it is reasonable to relate these two factors. All the models were only run for sleep duration, but the sleep data has other values like restless duration and restless count, and models could be run for these values as well. While the survey data was processed and mapped from survey response to subject ID, it was never used any further, so this data could be utilized. This survey data could be correlated with the model estimations, to further analyze the effectiveness of the different models. It would be possible to investigate modelling techniques other than GLM and GLMM and see their effectiveness, even though GLM and GLMM should in theory be the best choices for the given scenario. With all of this accomplished, a research paper could be written discussing the value of sensor-based sleep estimation, and whether or not any conclusions from the models can be used to help improve sleep quality.